SWITCHING APPARATUS FOR HIGH SPEED CHANNELS USING MULTIPLE PARALLEL LOWER SPEED CHANNELS WHILE MAINTAINING DATA RATE

Introduction

The present application is directed to switching apparatus for high speed channels using multiple parallel lower speed channels. Specifically, one application is in a switching network.

Background of the Invention

In a switching network, all receiving channels (or ports) route data to a switching fabric, which then switches the data, which is normally in the form of data packets of uniform or variable length, to a specific destination transmit channel (or port). Because fiber optic technology can support data rates much higher than traditional electrical standards, fiber optic channels have become the high-speed channel standard. Because the data rate of a single channel of a switching network is now likely to be higher than the data rate of a single fabric connection, multiple fabric connections must be used to support the data rate of the single channel.

Thus the prior art implemented multiple connections in parallel to increase the effective bandwidth of a single fabric connection. Fig. 1 illustrates this concept which is known as packet striping (or bit splicing) where the input channel or sender node is divided into several lower speed channels and then resequenced again at the receiver node. Thus a typical data packet is divided into parts or stripes with each part being sent on a separate fabric connection. With the four connections, the effective bandwidth of the overall fabric connection is increased by a factor of four even though the actual bandwidth of each connection is one-fourth of that.

In packet striping, as implemented in a typical switching fabric, the packet is divided into equal chunks (a chunk being a portion of a packet) and each chunk is sent to a separate switching plane of the switching fabric.

In a practical example, a data packet which is 40 bytes in length sent over a fabric consisting of four parallel paths must be divided into four 10-byte chunks. Since each chunk or packet portion requires its own so-called header for identifying that chunk and its origin and

destination, and this typically might require 2 bytes of information, this means that each transmitted data chunk has an overhead which is a substantial portion of the total data chunk. This effectively reduces the bandwidth by this amount (or, in other words, the effective data rate).

Object and Summary of Invention

It is therefore the general object of the present invention to provide switching apparatus for high speed channels using multiple parallel lower speed channels but maintaining a high data rate.

In accordance with the above object, there is provided a switching apparatus operating at a significantly higher data rate than switching elements(SEs) which form a switching fabric and operate at a lower data rate, the SEs routing data from at least one ingress source port, which receives data at the higher data rate, to egress destination ports the data being grouped in data packets having a uniform or variable plurality of digital bytes. The apparatus comprises an ingress source port including means for receiving successive data packets at the higher rate and for transmitting data via a plurality of output ports at the lower rate to the SEs. A sequential array of low data rate SEs each having a plurality of input ports are individually connected to each the output port of the source port the SEs and source port including means for switching source output ports successively from one SE to another available SE in response to a data packet event whereby the effective data rate from said source outputs to the SEs is at the higher data rate.

A method similar to the foregoing is also provided.

Description of the Drawings

Fig. 1 is a block diagram illustrating a prior art packet striping technique.

Fig. 2 is a representation of a data packet used in the present invention.

Fig. 3 is a simplified block diagram of switching apparatus incorporating the present invention.

Fig. 4 is a detailed showing a portion of Fig. 3 illustrating its operation.

Fig. 5 is a block diagram of Fig. 3 in greater detail.

Fig. 6 is a diagram useful in understanding the operation of Fig. 5.

Fig. 7 is a more detailed block diagram of a portion of Fig. 5.

Fig. 8 is a flow chart useful in understanding the operation of Fig. 7.

Fig. 9 is a flow chart useful in understanding the operation of the invention.

Fig. 10 is a block diagram illustrating an expandable variation of Fig. 3.

Detailed Description of Preferred Embodiments

Fig. 2 illustrates a typical data packet configuration which the switching apparatus of the present invention operates on. This data packet itself may consist of 40 or fewer digital bytes or up to 9,000. Attached to the data packet in a manner well known in the art is , for example, an 8-byte so-called header which contains priority, source and destination information.

Fig. 3 is an overall diagram of the switching apparatus where there are a number of ingress source ports 10 numbered 0 through 64 each receiving from, for example a framer which normally puts together a digital data packet, at a rate of 10 Gbps. The ingress ports 10 include a TM (traffic manager) and a communications processor (CP) and are labeled TM/CP. Each source port has an 8-line output port, each individually coupled to an input port of switch elements SE0 through SE7 which together create a so-called switching fabric. In turn, the eight switch elements each with 64 input ports and 64 output ports are similarly connected on an output side to egress ports 12 also designated CP/TM which have 8-line inputs and are numbered 0 through 63. The combination of the 64 ingress ports and 64 egress ports make up a 64 port full duplex port.

Again, as on the input side, each output port of a switch element has a direct serial link to one of the CP/TMs or egress port units. Then the egress ports 12 are coupled into, for example, a

high speed channel network (e.g., fiber optic) to transmit data at a 10 Gbps rate in a manner similar to the incoming data, but with the data having been rerouted to a selected destination port. Finally, as indicated in Fig. 3, the high input and output data rates of 10 Gbps cannot normally be sustained by the switch elements SE0 through SE7 which as indicated are limited to a lower data rate of 2.5 Gbps. Thus, in this practical embodiment the ratio of the higher data rate to the lower data rate is a 4:1 ratio.

Fig. 4 illustrates in very brief form the operation of the present invention where the ingress port 10 designated CP/TM0 receives data at the high data rate of 10 Gbps and then via a plurality of output ports distributes this input data on the eight lines 18, one line to each SE_0 through SE_7, at a lower data rate of 2.5 Gbps. Thus on each of the lines 18 a data packet is sent, for example, as indicated to switching element SE_7, and routed to a predetermined destination port 12. Thereafter in a sequential successive or round robin manner the next link 18 is used to transmit another data packet to SE_6 and then to SE_5. As indicated at 21, these are blocks of data versus a time axis. Some latency is present but this is a minimal tradeoff to achieve a greater throughput. In other words, over a single switching fabric multiple parallel lower speed channels are provided but the effective throughput of data is at the higher data rate and with a complete data packet being transmitted through one serial link.

Fig. 5 shows the input port arrangement 10 in greater detail. Here each communications processor CP0 through CP 64 is input linked to a framer 32 which, as discussed above puts together frames or packets. On the line 33 these are transferred to the communications processors and then to the traffic managers (TM) 34. The general functions of such traffic managers are to formulate an additional header for data packets to provide parsing, classification and editings; the traffic manager also determines to which switching element SE the data packet is to be transferred and to which port of that switching element. This is done in conjunction with the sequential sprinkler engine 35 (SSE) which is a part of each traffic manager. The output of the traffic manager is actually the output port lines 18 (see Fig. 5) of the ingress port 10. There is one line to each switching element SE_0 through SE_7. The output side of the switching apparatus, as also indicated in Fig. 3, is a duplicate with CP/TM 0 through 63 forming destination ports 12.

Sequential sprinkler engines 35 of each ingress port function in conjunction with a controller 38 and its table of destinations 39 to successively switch data packets from one source output port to another on the lines 18.

Each SSE 35 has its own controller and associated units. In operation the table of destinations 39 includes the last SE which has been used; to which a data packet has been transferred. Then in combination with the SSE 35 and controller 38 and under the control indicated by the function block 41, a switch occurs successively from one SE to another at each event to the next available SE. And this event is when another data packet is received by the traffic manager. Thus, the SSE 35 in effect "sprinkles" or distributes on a sequential or successive basis data packets from one SE to another in a manner that the high speed data rate is maintained while at the same time not utilizing in effect a single serial link for each data packet and avoiding the split up data into smaller units where overhead becomes a problem.

Fig. 6 illustrates in greater detail how the SSE 35 operates. Here from the traffic manager, indicated as being a FIFO (first in, first out memory), a line of data packets designated 1, 2 and 3 are being received. The first data packet is indicated as being sent to switching element 7. After this operation has started, a short time later, indicated as t_1, data packet 2 is transmitted (at the lower 2.5 Gbps rate) to SE_6. Then for data packet 3, at a later time t_2, its transmission to SE_6 is started. Due to the successive switching arrangement there is a latency but this is a minimal tradeoff to achieve greater throughput. As indicated by the logic unit 41 the availability of the SE may depend on whether it is being utilized at the moment for a previous data packet or has failed.

And, in fact, this illustrates the redundancy of the present invention where assuming an SE has failed, the logic assumes that this failed SE is busy and automatically goes to another switch element. For example, as illustrated in Fig. 3 with a 4:1 data input switch element data ratio, theoretically only four switch elements of the type illustrated are necessary. However, to provide for additional overhead due to headers, etc. additional bandwidth is provided by another two switch elements. In addition, to provide redundancy in case of failure of one of the SEs, two additional elements are provided. However, theoretically the number of switch elements may be exactly proportional to the ratio of data rates between the input data rate and the data rate capability of the switch elements. But throughput is still doubled even if only two switch elements are used. This may be feasible in some situations where there's not a constant high rate of data input.

Fig. 7 illustrates in greater detail one possible configuration of a traffic manager 34. Because of the high rate of data input through the traffic manager from the communications

processor (CP) and a slight time delay as illustrated in Fig. 6, some buffering must be included in the system. This is provided by a caching scheme. Such scheme is indicated in greater detail in a copending application, entitled Head and Tail Caching Scheme, Attorney Docket No. 6979/12. Referring in detail to Fig. 7, from the communications processor high speed data is coupled through the tail FIFO memory 41 and a multiplexer 42 to the head FIFO memory 43. Data packets will queue up as indicated in Fig. 6 as 1, 2 and 3 and be distributed by the sequential sprinkler engine (SSE) 35 and the read pointer (RP) to the various SEs as discussed. If data comes in at a rate faster than read or outputted to the switching elements fast, and the head FIFO memory 43 fills and the input data will start filling the tail FIFO memory 41. The write pointers and read pointers handle this detail under the control of memory controller 44 which has the WP and RP outputs. It is also coupled to the multiplexer 42. The tail or buffer FIFO 41 will initially keep the head FIFO memory 43 full as it is so-called de-queued (that is as it distributes data packets to the various switching elements). However, if the tail FIFO memory itself becomes full, then the so-called large scale off chip buffer memory 46 is utilized. Here as discussed in the above copending application uniform blocks of data on line 47 are transferred into the memory 46. And the transfer is arranged to be very efficient by use of uniform data block sizes. Finally, when the sudden burst of data packets decreases the traffic manager can de-queue all data from the large scale memory 46 and return to its normal functioning.

The above process is illustrated in Fig. 8 where in step 51 the head FIFO memory is first filled and then in step 52 the tail FIFO memory after the head FIFO overflows. And finally in step 53 the data is stored in the buffer memory until the tail FIFO has space. Then the data is retrieved to the tail buffer and finally written to the head FIFO.

As illustrated in Fig. 4, because of the asynchronous nature of the data inputs to the switching elements and its output as indicated by the time axis reordering may be necessary of the data. In other words, the present invention trades some sacrifices some latency to maintain the highest data rate throughput and enable simple redundancy. Referring to Fig. 9, one reordering technique is illustrated in flow chart form. Here in step 51 each data packet gets a time stamp when it leaves a source communication processor. Then on the output side, when the packets are received by the destination communications processor, they are put into a queue. Each destination CP has a separate queue. As the packets are received, the lowest time stamp is determined at step 53. A time out period occurs when this system clock reaches the value of the

lowest time stamp added to the minimum delay. If this time out period has not yet been released, the system repeats itself as illustrated in step 54. If it has occurred, as shown in step 55, it is now theoretically known that all frames have been received (assuming no other problems) and the packet with the lowest time stamp is placed at the head of the queue. This is just one illustration of reordering and others may be used. However, details of the reordering technique may be found in a copending application titled "Reordering of Sequence Based Packets in a Switching Network:," Attorney Docket #06979/08.

To provide additional data ports, the switching fabric of the switching elements shown in Fig. 3 is easily scalable or expandable to accommodate greater data input. One technique is a butterfly expansion, illustrated in Fig. 10. Here there are the original SEs, SEO and SEI are so labeled. To expand additional switching elements designated SE2' – SE5' are connected with the designated interconnections that double the amount of input and output ports.

To summarize the operation of the invention, a uniform or variable length data packet is stored in an ingress port at a relatively high data rate and is transmitted to its final destination port on one serial link. Moreover, since the packet is not broken into smaller pieces , where the header becomes a significant part of the data packet, overhead is minimized and the highest data rate is maintained. The switching fabric configuration as shown by the switch elements of Fig. 3 allows for redundancy where, in the case of failure one switch element, another is automatically selected. This is not true of ordinary parallel channel devices as illustrated in Fig. 1. Moreover, additional bandwidth and data input can be provided by adding more switch elements; for example, in a butterfly configuration as illustrated in Fig. 10.

In summary, improved switching apparatus for increasing data rates with limited switching speeds has been provided.